Machine learning from crowds
A systematic review of its applications

Enrique G. Rodrigo [*], Juan A. Aledo[†], José A. Gámez[‡]

Abstract

Crowdsourcing opens the door to solving a wide variety of problems that previously were unfeasible in the field of machine learning, allowing us to obtain relatively low cost labeled data in a small amount of time. However, due to the uncertain quality of labelers, the data to deal with is sometimes unreliable, forcing practitioners to collect information redundantly, which poses new challenges in the field. Despite these difficulties, many applications of machine learning from crowdsourced data have recently been published that achieve state of the art results in relevant problems. We have analyzed these applications following a systematic methodology, classifying them into different fields of study, highlighting several of their characteristics and showing the recent interest in the use of crowdsourcing for machine learning. We also identify several exciting research lines based on the problems that remain unsolved to foster future research in this field.

This is the accepted version of:

Enrique González Rodrigo, Juan A. Aledo, Jose A. Gámez.
Machine learning from crowds: A systematic review of its applications.
Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 9(2) (2019)
https://doi.org/10.1002/widm.1288

Please, visit the provided url to obtain the published version.

[*]Department of Computer Systems, University of Castilla-La Mancha
[†]Department of Mathematics, University of Castilla-La Mancha
[‡]Department of Computer Systems, University of Castilla-La Mancha

1 Introduction

With the recent appearance of crowdsourcing platforms such as *Amazon Mechanical Turk* a great number of machine learning practitioners have expressed interest in using them to increment the efficiency and scope of their work. Several problems that would be too expensive to deal with using traditional methods now become easier, while problems which were not feasible are now tractable. Citing Howe, the term *crowdsourcing* can be defined as (Howe, 2006):

> *"[...] the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call."*

For Howe, the use of crowdsourcing involves two clearly defined elements: a generally large group of people and an open call. Recently, though, in the area of machine learning this term has been used to refer not only to humans but to other elements, such as sensors and algorithms. Even the open call requisite has been relaxed, allowing the use of a small group of people known *a priori*. In other words, when we talk about *crowdsourcing* problems in the context of this paper we refer to problems that use a group of *elements* that provide noisy data for a given example and whose quality we may, or may not, know much about. These elements could be, for instance, a group of experts examining medical images of patients, antivirus programs analyzing threats or Amazon Mechanical Turk workers analyzing facial expressions. These groups share one characteristic: there is some kind of uncertainty in the information provided by the members of the group, as some of them may be better at that task while others may have misunderstood the problem (or be malicious). The result of this annotation process is a dataset which is inherently noisy and that needs to be preprocessed to be useful in a machine learning task.

There are several applications of machine learning that meet the above description. For example, in the area of computer-aided diagnosis, we might want to determine whether a tumour in a medical image is benign or malign. However, in most cases, obtaining an accurate label to train is very costly, so normally a group of experts is asked to give their opinions about the image. Unfortunately, these professionals will probably have a different background, causing disagreements during the labeling process (Raykar et al., 2010). Another area of study where these algorithms could be of great help is the aesthetic image classification, in which one of the goals consists in creating a model able to distinguish between great and average images. However, the subjectiveness of the task makes it difficult to determine the ground truth necessary for standard models (Datta, Joshi, Li, & Wang, 2006). Despite of this, we could use crowdsourcing platforms to collect several potentially noisy labels that could allow us to build an accurate model taking into account this diversity of opinions.

In this work, we analyze several applications dealing with crowdsourced data, in order to show the growing interest in this field, as well as the techniques being used and the problems being faced by researchers using crowdsourcing to develope state of the art machine learning solutions. As proof of the interest in this topic, several reviews (e.g (J. Zhang, Wu, & Sheng, 2016) and (Zheng, Li, Li, Shan, & Cheng, 2017)) about machine learning techniques and tools in this area have been published recently. However, these reviews do not include recent applications of these techniques, leaving a gap that this work is intended to fill. This kind of review, as can be found in other fields (Paliwal & Kumar, 2009; B. Chen & Cheng, 2010; Rashid & Rehmani, 2016), can foster research, indicating challenges that could lead to new developments in the future.

The paper is organized as follows. In Section 2 we describe the systematic methodology followed, as well as the goals of this review. In Section 3 we analyze the research interest in applications of learning from crowdsourced data. In Sections 4 and 5 we analyzed the type of crowd and the machine learning techniques used in these applications, while in Section 6 we show the main areas of application of crowdsourced machine learning. In Section 7 we discuss the main problems that remain unsolved and that could lead to future research in the field. Finally, in Section 8 we present the conclusions of this work.

2 Methodology

This review follows the systematic procedure proposed in (Kitchenham, 2004). We start from various questions that are of interest and perform several steps with the goal of finding the relevant literature to augment our understanding of these questions, in a reproducible way. In this section, we summarize the details of the followed procedure.

2.1 The need for a review

As seen in the introduction, learning from the wisdom of crowds is an interesting topic for several reasons. First of all, it allows us to tackle problems without enough available ground truth data, but in which the use of a group of people is possible. It also allows the collection and use of labeled data for problems where there is not an objective ground truth, such as affective behaviour recognition (Nicolaou, Pavlovic, & Pantic, 2014). Generally, it also reduces the costs of the data gathering phase in supervised machine learning experiments (J. Zhang et al., 2016).

To our knowledge, there has not been any attempt to explore the applications of the techniques involved in learning from crowdsourced data. An effort in this direction could reveal new branches for future research in this field, as well as ways to improve the results

of applications through the use of more powerful algorithms. In fact, we can find several reviews related to, but not exploring, this specific issue: two reviews of algorithms for solving the problem of learning from crowdsourced labeled data (J. Zhang et al., 2016; Zheng et al., 2017); the challenges of using crowdsourcing from a system design perspective (Garcia-Molina, Joglekar, Marcus, Parameswaran, & Verroios, 2016); a taxonomy of crowdsourcing tasks not necessarily from a machine learning perspective (Good & Su, 2013); a survey of data management techniques with crowdsourced data (G. Li, Wang, Zheng, & Franklin, 2016); projects related to applying crowdsourcing for climate and atmospheric sciences (Muller et al., 2015); uses of crowdsourcing in the different phases of software engineering (Mao, Capra, Harman, & Jia, 2017); incentives used in the crowdsourcing literature (Gao et al., 2015); the concept of crowd intelligence, as well as the platforms and research problems associated with it (W. Li et al., 2017); and techniques related to task design, assignment and quality control (Chittilappilly, Chen, & Amer-Yahia, 2016). These publications show a great interest in the topic, not only for the opportunities that it offers for data scientists, but also for the practitioners in other areas of research which could benefit from these new techniques.

2.2 Review questions

In this review, we analyze applications of machine learning from data provided by crowds, in the form of both labels and features. Specifically, we address the following research questions:

Q1 What has been the interest in the topic in the last decade? [1]

Q2 Which areas are associated with the greatest interest in the topic?

Q3 How are the crowds used to achieve the goals of the application?

Q4 What are the most commonly used techniques for tackling the caveats of using crowdsourcing for machine learning?

Q5 What interesting future research lines follow from the applications?

We believe this information could be highly beneficial for machine learning practioners as well as for researchers interested in gaining knowledge about how to learn from crowdsourced data.

2.3 Search process

We used three methods to obtain the articles reviewed in this work:

[1] Although there are algorithms related to learning from crowds previous to 2010, the applications found date from 2010 on. Therefore, we have fixed 2010 as the first year of our study.

- Database search, using search strings related to the goals of this review.
- Citation analysis of the main machine learning algorithms present in the literature (J. Zhang et al., 2016; Zheng et al., 2017)

2.3.1 Database search

The following databases were used to gather articles containing applications of machine learning from crowdsourced data: Scopus [2], Web Of Science [3], Google Scholar [4], DBLP [5], ScienceDirect [6], ACM Digital Library [7], and IEEE Xplore Digital Library [8]. In each of these databases we used a search string which, in general, searched for the presence of two terms in the title of publications:

- Machine learning related terms, such as learning or classification.
- Crowdsourcing related terms, such as *crowds* or *annotators*.

Where possible, we reduced the results from thee search by only looking for articles and conference publications, and limiting the topics of both conferences and journals to those related to machine learning. For the specific strings for each database, we refer the reader to Appendix A.

As the search strings used were quite general (with the goal of obtaining a large number of publications) we then refined the results by reading the titles and abstracts within the articles, filtering out those that did not align with the topic of this paper. Specifically, all papers had to be about applications of machine learning from crowdsourced labeled data.

2.3.2 Citation analysis

In this phase, we took the algorithms analyzed in (J. Zhang et al., 2016; Zheng et al., 2017) and searched for their citations on the *Scopus* platform, carefully selecting the papers related to the subject of this review.

2.4 Quality assessment

All the works analyzed in this review were published in peer reviewed journals or conferences of recognized scientific value, which guarantees that they meet a certain standard of quality.

[2] https://www.scopus.com
[3] https://http://apps.webofknowledge.com
[4] https://scholar.google.es
[5] http://dblp.uni-trier.de
[6] http://www.sciencedirect.com
[7] http://dl.acm.org
[8] http://ieeexplore.ieee.org/Xplore/home.jsp

2.5 Data extraction

Once we established the pool of articles related to the topic of this survey, we went through them to answer the following questions:

Q1 What has been the interest in the topic in the last decade?

- Year of publication.
- Publication type: journal or conference.
- Author's country.
- Publication citations.

Q2 Which areas are associated with the greatest interest in the topic?

- Area of knowledge of the application

Q3 How are the crowds used to achieve the goals of the application?

- Information provided by the annotator (for example, labels or features)
- Platform used for the annotation process

Q4 Which are the most commonly used techniques for tackling the caveats of using crowdsourcing for machine learning?

- Learning algorithm family.

Q5 What interesting problems arise from the analyzed applications?

- New problems arising from each publication, if any.

3 Research interest analysis

We have found a total of 116 publications, 51 of which are journal publications, while 65 are published in conferences. To see the increasing interest in the research in this area, it can be observed the yearly increase in the number of publications in Figure 1, which indicates a rising interest in using crowdsourced data for machine learning applications.

It is also of interest to see how the publications are distributed geographically. For this, we took the country of the publications' authors[9] of the publications and made two graphs. The graph on the left (Figure 2a) shows the number of publications corresponding

[9]We counted the countries related to a publication once. For example, if a publication has three authors from two different countries, each country was counted once

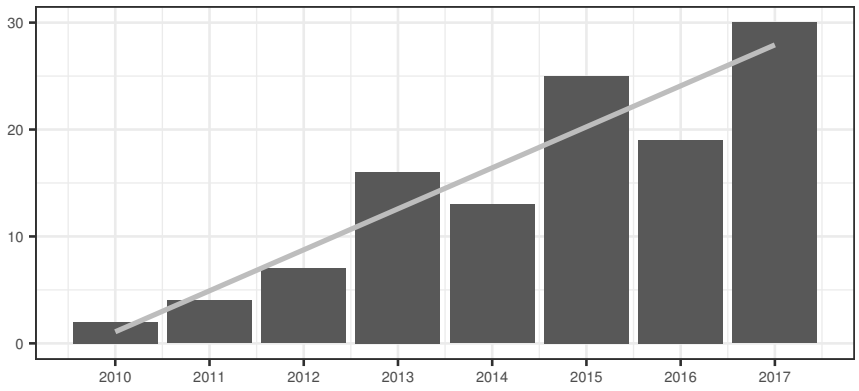
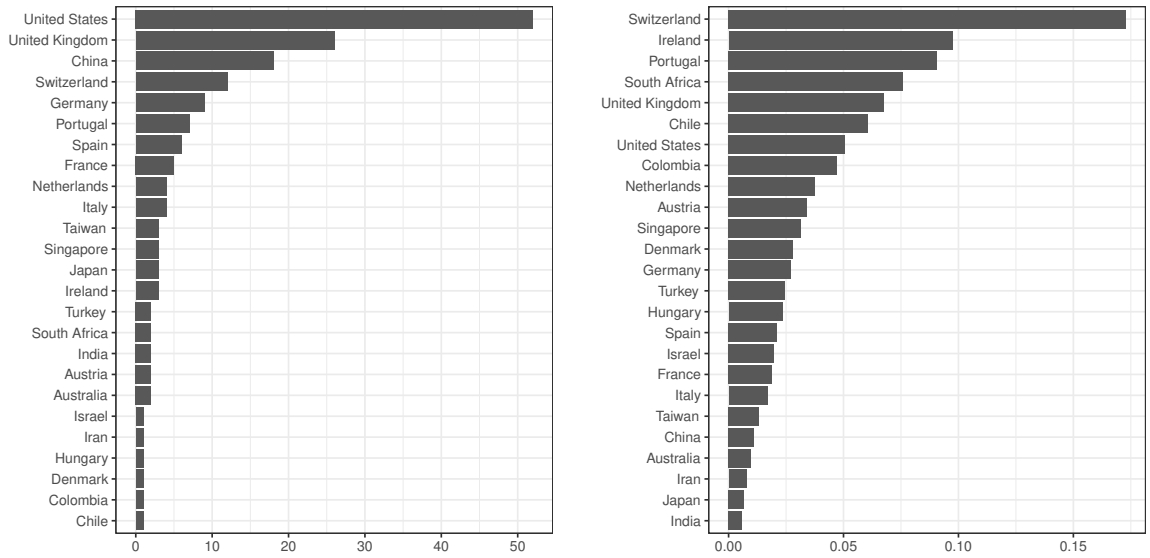


Figure 1: Number of publications by year

to each country. The graph on the right (Figure 2b) takes into account the total number of publications by country indexed by SJR (Elsevier, 2017) in the area of Artificial Intelligence, dividing the number of publications found in this study by the total number of publications for each country and then normalizing the results [10].



(a) Publications by country

(b) Publications by country (Ratio)

Figure 2: Number of publications by country in comparison with the total for the country.

Compared to the total number of publications for each country, it seems that there is a greater than expected number of publications in countries such as Switzerland or UK,

[10] We divide the publications by the total number given by SJR and then normalize, so that they all add up to one. Although the data from SJR is for journals covering only the period from 2010 to 2016, it allows us to compare countries with big differences in the total number of publications.

while countries such as China have fewer publications related to the applications of machine learning from crowdsourced data.

As another interest indicator, we analized the mean number citations for each work using the total number of citations for a publication obtained through Scopus, as this platform provides accurate citation information. The mean number of citations by year can be seen in Table 1. It is significant the high mean number of citations for the papers published three or more years ago.

2010	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**
22.5	19.75	23.26	28.25	14.62	4.56	4.89	1.73

Table 1: Mean number of citations by year of publication

4 Crowd use analysis

In this section we explore the way in which the crowd is used to tackle the different problems, in terms of:

- Information obtained from the crowds (labels, features,...)
- Characterization of *workers* (experts, algorithms...)
- Used platforms (Amazon Mechanical Turk, CrowdFlower,...)

4.1 Information obtained from the crowds

In our research, we have identified 3 different types of information obtained using crowdsourcing: labels (e.g. for supervised machine learning), features, and a mixture of features and labels (Both). The distribution of the results can be seen in Figure 3.

The predominant use of crowdsourcing for learning from crowdsourced data is the elicitation of labels for a dataset. This use is also the most studied in the literature and there exist several algorithms designed to improve the results when learning from them. However, there is also the need to obtain features, and also features and labels jointly, in order to create new datasets which may lead to interesting future opportunities.

4.2 Characterization of workers

Although the term *crowdsourcing* usually refers to the use of a generally large group of (possibly) unreliable people, we have identified other different types of groups used for the

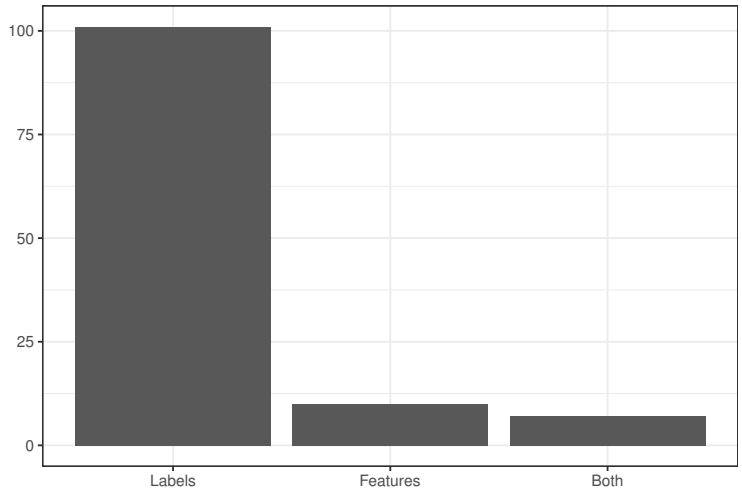


Figure 3: Distribution of type of crowd used in the publication

same purpose and using similar crowdsourcing techniques: experts, volunteers, algorithms, etc. These other types of groups, albeit with characteristics distinguishing them from the standard crowdsourcing definition, are usually treated with the same methods, and that is why they are included in this study. The distribution of the number of publications for each category can be seen in Figure 4. In this figure, by the term *crowd* we refer to the use of (generally) non expert people, via platforms such as Amazon Mechanical Turk or CrowdFlower, who may receive some incentive. This is the use case that is nearest to the classic definition of crowdsourcing. This is also the most common use case. However, the use of experts is also very important in this field, coming in second place [11]. We also distinguish between the above two categories and the one where the labeling process is performed by volunteers, since normally the characteristics of the problems solved by them fall between those of the two categories mentioned above. [12]. Apart from the use of people, we have also identified the use of other elements, such as algorithms or sensors. In the applications, these elements are treated with methods similar to the ones used with people. However, the number of publications using them is small with respect to the ones using people as annotators.

[11] As we show later, this is due to use cases where a group of experts is usually preferred, as in applications related to medicine

[12] In the applications reviewed, volunteers usually exhibit more willingness to work accurately than paid crowd workers, although, in general, they are not as accurate as experts in dealing with some of the tasks

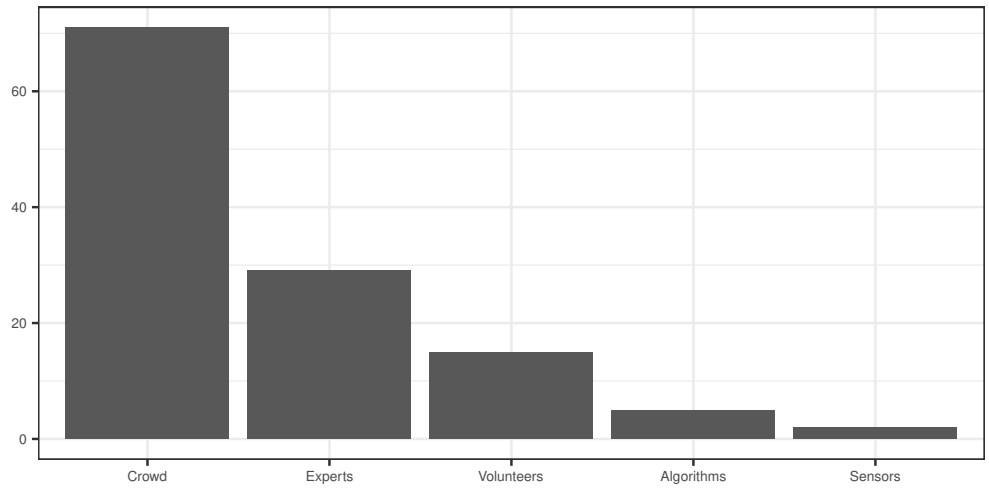


Figure 4: Distribution of publications by worker type

4.3 Platforms used

In Figure 5, we show the crowdsourcing platform used in the publications from the *Crowd* category in the above analysis, that is, the publications that use platforms available to the general public.

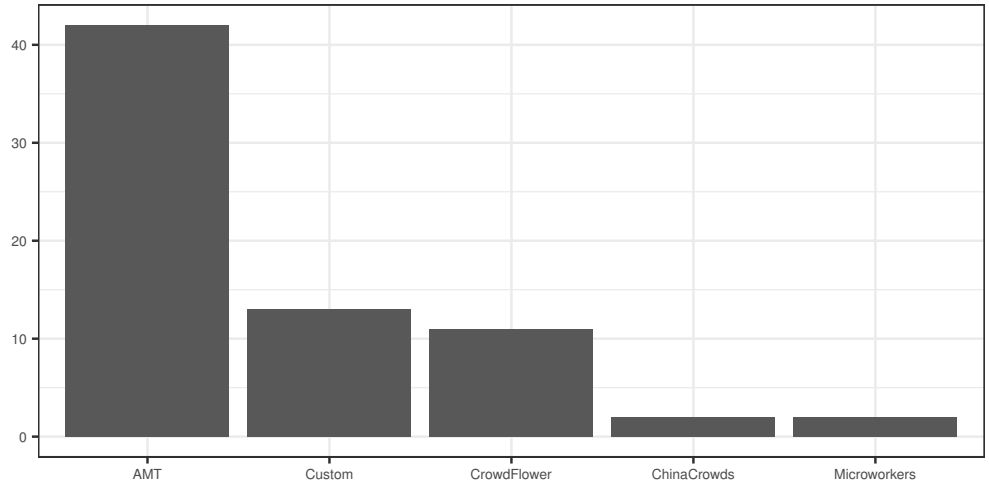


Figure 5: Distribution of *crowd* category publications by platform

As may be expected, Amazon Mechanical Turk [13] was the most commonly used platform in the applications. Surprisingly, the second one was the author's own custom solution for

[13] https://www.mturk.com

the problem, followed by CrowdFlower[14], Microworkers[15] and ChinaCrowds[16].

5 Technique analysis

Normally, when learning from crowds, practitioners try to aggregate crowd responses so that the data is more accurate. We have grouped the publications into 3 categories: those using simple aggregation algorithms (Simple), such as the mean or majority voting of the annotations (most frequent value), those using complex algorithms for aggregating (Complex), such as algorithms learning annotators reliability iteratively[17], and those applications where no aggregation of data was used (No). Figure 6 shows the distribution of the publications in these categories.

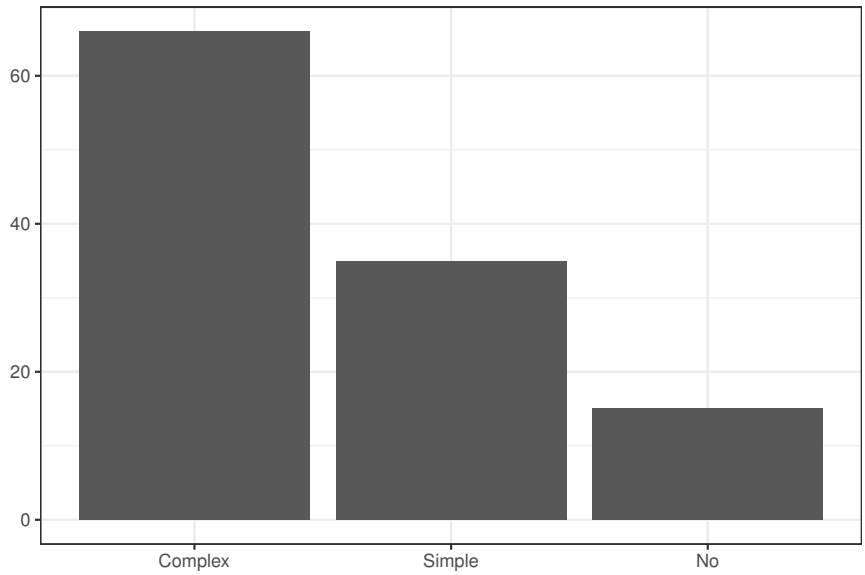


Figure 6: Number of publications by aggregation type

As can be seen in Figure 6, although the use of complex aggregation is the most common approach, the use of simple aggregation is also very popular. On the other hand, more than 10 publications use the labels without any aggregation, as if they were the true labels themselves.

Inside de Complex aggregation category, we find a great number of different algorithms, most of them particularly designed to solve the task of aggregation of crowdsourced data for the precise problem they are trying to solve. However, following the taxonomy from

[14] https://www.crowdflower.com
[15] https://microworkers.com
[16] http://www.chinacrowds.com
[17] For a recent comparison of these algorithms see (Zheng et al., 2017)

(Zheng et al., 2017), we can further divide the algorithms regarding the technique used, either optimization, i.e. algorithms capture the relations between workers and tasks using an optimization function, or PGMs, i.e. algorithm designed using a probabilistic graphical model. In Figure 7, one can see that, although optimization is also used in applications, the majority of the algorithms use an approach based on PGMs.

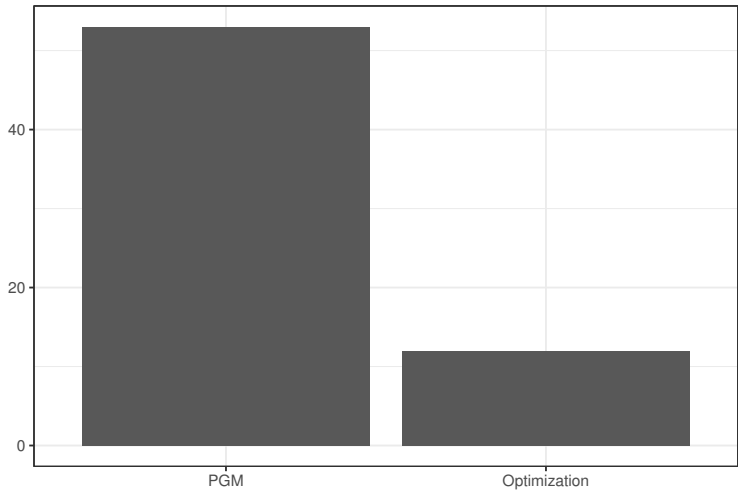


Figure 7: Technique used in publications using complex aggregation

6 Publication areas

Figure 8 shows the distribution of applications by knowledge area[18]. Areas such as *Bioinformatics*, *Computer Vision* and *Natural Language Processing* are the areas where these techniques are most frequently applied, as some of the tasks align perfectly with the problems that crowdsourcing is trying to solve.

In Figure 9a, we show the type of crowd (see Section 4) used in each area. Bioinformatics is the field with the greatest number of publications related to crowdsourced machine learning, a great number of them involve aggregating expert opinions in problems where the ground truth is not known (or difficult to know) and in which the use of non experts may be impossible due to the difficulty of the task. Other fields such as Natural Language Processing or Computer Vision share a similar distribution, as the most common use of crowdsourcing is through the crowdsourcing platforms. In the rest of applications the use of crowds is the most common use case, although the other approaches are also used.

[18] Only one area for publication was extracted, taking into account the topic of the article and the journal in which it was published.

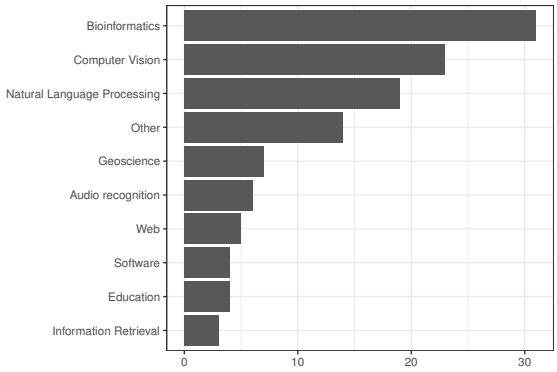
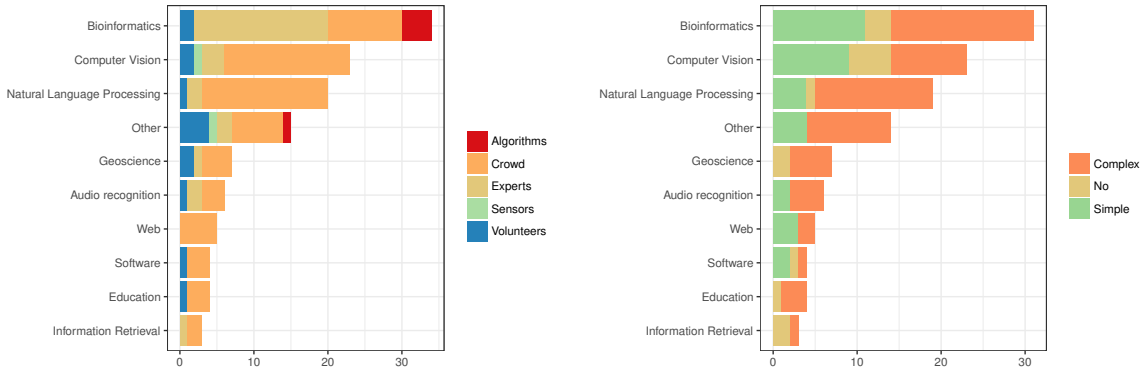


Figure 8: Distribution of publications by knowledge area



(a) By worker type

(b) By technique

Figure 9: Knowledge area by worker type and technique

We can also compare which kind of aggregation is performed by area (Figure 9b). In this case, we find that areas such as Bioinformatics or Natural Language Processing seem to prefer aggregating results using complex algorithms, while other areas, such as Computer Vision, or Software tend to use a simpler approach.

Next we examine each publication, according to these areas of knowledge, in more detail.

6.1 Bioinformatics

There is a great interest in crowdsourcing techniques in the Bioinformatics community. Although the majority of publications use a small group of experts as annotators, several applications use platforms such as Amazon Mechanical Turk to obtain labels. Moreover, some applications use a mixture of algorithms and experts, as well as volunteers. In Table 2 we show the references of the applications found, as well as information about the type of

Type	Application	Aggregation	Crowd	Platform
Medical Images	(Kaster et al., 2010)	Complex	Experts	Custom
	(Luengo-Oroz et al., 2012)	Simple	Crowd	Custom
	(Mavandadi et al., 2012)	Complex	Experts	Custom
	(DeFelipe et al., 2013)	Simple	Experts	Custom
	(Mitry et al., 2013)	Simple	Crowd	AMT
	(Chatelain et al., 2013)	Complex	Expert	Custom
	(Mahapatra et al., 2014)	Complex	Experts	Custom
	(Mihaljević et al., 2015)	Simple	Experts	Custom
	(Kaya & Can, 2015)	Simple	Experts	Custom
	(Ataer-Cansizoglu et al., 2015)	No	Experts	Custom
	(Albarqouni et al., 2016)	Complex	Crowd	CloudFlower
	(Sameki et al., 2016)	No	Crowd	AMT
	(V. Chang et al., 2017)	Simple	Experts	Custom
	(Sharma et al., 2017)	Simple	Crowd	CrowdFlower
	(Brady et al., 2017)	Complex	Labels	AMT
Biomedical Information Extraction	(Greenwood et al., 2013)	Simple	Experts,Crowd	AMT,Custom
	(Tastan et al., 2014)	Complex	Experts	Custom
	(Khare et al., 2015)	Complex	Crowd	AMT
	(Jain et al., 2016)	Complex	Algorithms	Custom
	(de Herrera et al., 2016)	No	Crowd	CrowdFlower
	(Wallace et al., 2017)	Simple	Experts	Custom
	(Ma et al., 2017)	Complex	Volunteers	Custom
Others	(Lu et al., 2011)	Complex	Algorithms,Experts	Custom
	(Silva et al., 2013)	Complex	Algorithms,Experts	Custom
	(Peng et al., 2013)	Complex	Experts	Custom
	(Zhu et al., 2014)	Complex	Experts	Custom
	(Zhu, Dunkley, et al., 2015)	Complex	Experts	Custom
	(Tan et al., 2015)	Simple	Volunteers	Custom
	(González et al., 2015)	Complex	Experts	Custom
	(Zhu, Pimentel, et al., 2015)	Complex	Algorithms	Custom
	(Lou et al., 2017)	Complex	Crowd	AMT

Table 2: Publications in Bioinformatics

aggregation, the type of crowd and the platform used[19] sorted by year.

Specifically, the area of medical images has recently expressed interest in this field as numerous applications require a great effort for labelling images (de Bruijne, 2016; S. Wang & Summers, 2012), such as image segmentation or cell classification. In this way, the majority of the applications of learning from crowdsourced data in this field take advantage of these techniques for reducing the labelling effort as well as improving results in fields where labelling is not trivial. In this sense, we have found applications about: tumor segmentation (Kaster et al., 2010); remote malaria diagnosis (Mavandadi et al., 2012) and malaria parasite quantification (Luengo-Oroz et al., 2012); classifying GABAergic interneurons (DeFelipe et al., 2013; Mihaljević et al., 2015); midbrain 3D ultrasound image segmentation (Chatelain et al., 2013); Crohn's disease segmentation (Mahapatra et al., 2014); retinal fundus classifi-

[19]If in the application the authors do not use any public platform (as is the case, normally, when using a group of experts) the term Custom is used in that column

cation (Mitry et al., 2013; Brady et al., 2017); predicting malignancy of pulmonary nodules (Kaya & Can, 2015); prematurity diagnosis in retinopathy (Ataer-Cansizoglu et al., 2015); mitosis detection in breast cancer (Albarqouni et al., 2016); melanoma cell segmentation (Sameki et al., 2016); sperm analysis (V. Chang et al., 2017) and segmentation of chromosomes (Sharma et al., 2017).

There is also interest in the biomedical information extraction community, which becomes clear from the fast growing of the number of related publications and experimental data. Although there are several applications and techniques for extracting information (Fleuren & Alkema, 2015), several problems do need information that only a human or even an expert can provide as the reader can find in the following applications: extraction of patient's personal experiences (Greenwood et al., 2013); refining curated protein interactions (Tastan et al., 2014); drug indication curation (Khare et al., 2015); extraction of information such as phenotype or stage of an study (Jain et al., 2016); biomedical compound figure annotation for publications (de Herrera et al., 2016); identifying reports of randomized controlled trials (Wallace et al., 2017) and drug side-effects discovery (Ma et al., 2017).

Other interesting applications in this area include: estimation of respiratory rate from the gene normalization (Lu et al., 2011); estimation of fetal heart rate, interbeat intervals and fetal QT intervals with noninvasive ECG (Silva et al., 2013); protein folding (Peng et al., 2013); ECG signal classification (Zhu et al., 2014; Zhu, Dunkley, et al., 2015); photoplethysmograms (Zhu, Pimentel, et al., 2015); sleep spindle detection (Tan et al., 2015); assessment of voice pathologies (González et al., 2015) and learning for ICD-11 sanctioning rules (Lou et al., 2017).

6.2 Computer vision

In computer vision, there is also a large number of publications covering different topics. However, unlike Bioinformatics, in this case most of the applications use crowd platforms such as Amazon Mechanical Turk. A summary of the applications seen for this area can be found in Table 3.

Type	Application	Aggregation	Crowd	Platform
Affective Interaction	(Wan & Aggarwal, 2014)	Complex	Experts	Custom
	(Nicolaou et al., 2014)	Complex	Experts	Custom
	(Katsimerou et al., 2016)	No	Crowd	Microworkers
	(Tavares et al., 2016)	Complex	Crowd	AMT
Object recognition	(Su et al., 2012)	No	Crowd	AMT
	(Salek et al., 2013)	Complex	Crowd	AMT
	(Vijayanarasimhan & Grauman, 2014)	Simple	Crowd	AMT
	(Bernaschina et al., 2014)	Simple	Crowd	Custom
	(Cabezas et al., 2015)	Simple	Crowd	Microworkers
Activity recognition	(Nguyen-Dinh et al., 2013)	Simple	Crowd	AMT
	(Nguyen-Dinh et al., 2014)	Simple	Crowd	AMT
	(Nazábal et al., 2016)	Complex	Sensors	Custom
	(Kratz & Wiese, 2016)	No	Crowd	AMT
Others	(Chittaranjan et al., 2011)	Complex	Experts	Custom
	(Srivastava et al., 2013)	Simple	Volunteers	Custom
	(Rudinac et al., 2013)	No	Crowd	AMT
	(Wu et al., 2013)	No	Crowd	AMT
	(Oosterman et al., 2015)	Simple	Crowd	CrowdFlower
	(Baklanov et al., 2016)	Complex	Crowd	Custom
	(Y.-L. Fang et al., 2017)	Complex	Crowd	Crowdflower
	(Servajean et al., 2017)	Complex	Crowd	AMT

Table 3: Publications in Computer Vision

In the area of affective interaction and emotion recognition (for an introduction to the topic, see (Kołakowska, Landowska, Szwoch, Szwoch, & Wrobel, 2014)) where there is a certain subjective component, crowdsourcing has a special relevance, as several algorithms allow to take into account the capabilities of an annotator for labeling certain types of cases. In particular, we have found applications related to: spontaneous facial expression recognition (Wan & Aggarwal, 2014); affective behaviour recognition (Nicolaou et al., 2014); emotion and mood recognition (Katsimerou et al., 2016) and facial expression classification for affective interaction (Tavares et al., 2016).

In the topic of object recognition (the reader may find an introduction in (X. Zhang, Yang, Han, Wang, & Gao, 2013)), the crowd is used for segmenting and labelling images. The goal here is not to solve problems with subjective components but taking advantage of the crowd to quickly process images that would be used in a machine learning process. We have found the following works related to this problem: obtaining regions of interest from an image (Su et al., 2012; Cabezas et al., 2015); image object localization (Salek et al., 2013); labeling crawled data for object detection (Vijayanarasimhan & Grauman, 2014) and using games for segmenting images (Bernaschina et al., 2014).

The previous goal is also shared in the field of gestures (L. Chen, Hoey, Nugent, Cook, & Yu, 2012) and activity recognition (Aggarwal & Ryoo, 2011), where crowdsourcing also becomes a powerful tool: human activity tagging (Nguyen-Dinh et al., 2013); online gesture recognition (Nguyen-Dinh et al., 2014); daily human activity recognition (Nazábal et al.,

Application	Aggregation	Crowd	Platform
(Costa et al., 2011)	Simple	Crowd	Custom
(Ng & Kan, 2012)	Simple	Crowd	CrowdFlower
(Passonneau et al., 2012)	Complex	Experts,Crowd	AMT
(Jones, 2012)	No	Crowd	AMT
(Machedon et al., 2013)	Simple	Crowd	AMT
(Salter-Townshend & Murphy, 2013)	Complex	Crowd	AMT
(Fornaciari & Poesio, 2014)	Complex	Crowd	Custom
(Rodrigues et al., 2014)	Complex	Crowd	AMT
(Hovy et al., 2014)	Complex	Crowd	AMT
(Huang et al., 2015)	Complex	Crowd	AMT
(Duan et al., 2015)	Complex	Crowd	Custom
(R. Yan et al., 2015)	Complex	Crowd	AMT
(Zhou et al., 2017)	Complex	Features	AMT
(Q. V. H. Nguyen et al., 2017)	Complex	Crowd	AMT
(Rodrigues et al., 2017)	Complex	Crowd	AMT
(A. T. Nguyen et al., 2017)	Complex	Crowd	AMT
(Z.-X. Li et al., 2017)	Complex	Crowd	AMT

Table 4: Publications in Natural Language Processing

2016) and gesture segmentation (Kratz & Wiese, 2016).

Other interesting applications include: detecting the most dominant person of the group with audiovisual features of group interaction (Chittaranjan et al., 2011); YouTube video categorization (Srivastava et al., 2013); learning user preferences for visual summarization (Rudinac et al., 2013); a framework for multimedia quality of experience evaluation (Wu et al., 2013); labeling visual artworks (Oosterman et al., 2015); cropland image classification (Baklanov et al., 2016); bumblebee image classification (Siddharthan et al., 2016); dog breed recognition (Y.-L. Fang et al., 2017) and plant type classification (Servajean et al., 2017).

6.3 Natural Language Processing

As was the case with computer vision, most of the publications use crowd platforms such as Amazon Mechanical Turk to obtain useful information for their problems. The list of applications found in this area can be seen in Table 4.

The publications found are diverse, with both problems where subjectivity is involved and problems where the main advantage of using crowdsourcing is the capability of providing inexpensive data. For an introduction to some of the following topics, we refer the reader to (Hirschberg & Manning, 2015). Especifically, we have found applications about: sentiment analysis of online media (Salter-Townshend & Murphy, 2013; Brew, Greene, & Cunningham, 2010); joke's humour classification (Costa et al., 2011); temporal relation classification (Ng & Kan, 2012); word sense (Passonneau et al., 2012); marketing messaging classification on Twitter (Machedon et al., 2013); POS tagging (Hovy et al., 2014); identifying fake Amazon reviews (Fornaciari & Poesio, 2014); sequence labeling (Rodrigues et al., 2014; A. T. Nguyen

et al., 2017); estimation of discourse segmentation (Huang et al., 2015); emotion estimation from narratives (Duan et al., 2015); crowdsourced translation (R. Yan et al., 2015); entity disambiguation (Zhou et al., 2017; Q. V. H. Nguyen et al., 2017; Z.-X. Li et al., 2017) and topic models (Rodrigues et al., 2017).

6.4 Geoscience

We have also found several applications that use crowds to classify land images and detecting events for geographical applications (Table 5): land image classification (Pistorius & Poona, 2014; Jia et al., 2016; Chesnokova, Nowak, & Purves, 2017); attribute mapping (Foody et al., 2015); human settlement mapping (Gueguen et al., 2017) and detecting geographical events (Garcia-Ulloa, Xiong, & Sunderam, 2017).

Type	Application	Aggregation	Crowd	Platform
Geoscience	(Pistorius & Poona, 2014)	No	Volunteers	Custom
	(Foody et al., 2015)	Complex	Crowd	Custom
	(Jia et al., 2016)	Complex	Experts	Custom
	(Gueguen et al., 2017)	Complex	Crowd	Custom
	(Chesnokova et al., 2017)	No	Crowd	Custom
	(Garcia-Ulloa et al., 2017)	Complex	Volunteers	Custom
Audio Recognition	(Ni et al., 2013)	Complex	Experts	Custom
	(Hantke et al., 2016)	Simple	Crowd	CrowdFlower
	(Tu et al., 2016)	Complex	Experts	Custom
	(Hantke et al., 2017)	Complex	Labels	Custom
	(S. Zhang et al., 2017)	Simple	Labels	AMT
	(Chapaneri & Jayaswal, 2017)	Complex	Crowd	AMT
Web	(Crescenzi et al., 2013)	Complex	Crowd	AMT
	(S. Chang et al., 2015)	Simple	Crowd	AMT
	(Min et al., 2017)	Simple	Crowd	AMT
	(Mok et al., 2017)	Simple	Crowd	AMT,CrowdFlower
	(Tacchini et al., 2017)	Complex	Crowd	Custom
Software	(Kong et al., 2015)	Complex	Crowd	Custom
	(Davami & Sukthankar, 2015)	Simple	Crowd	Custom
	(Nazar et al., 2016)	Simple	Volunteers	Custom
	(J. Wang et al., 2017)	No	Both	Custom

Table 5: Publications in geoscience, web, audio recognition and software

6.5 Audio Recognition

In the area of audio recognition (Table 5), we found publications looking for labeling databases for several problems, not only for speech (Hantke et al., 2016; Tu et al., 2016) but for music (Ni et al., 2013; Chapaneri & Jayaswal, 2017) and emotion recognition (Hantke et al., 2017). There is also one application dealing with acoustic classification of animal species (S. Zhang et al., 2017).

6.6 Web

Different Web related problems (Table 5) can also benefit from using crowdsourcing techniques, especially in processing social generated content, where we found applications dealing with social media posts (S. Chang et al., 2015), fake content (Min et al., 2017; Tacchini et al., 2017) or the quality of the reviews (Mok et al., 2017), and on wrapper generation (Crescenzi et al., 2013).

6.7 Software

In the area of software development (Table 5), we found publications related to various stages of the development process: improving performance of applications using crowdsourcing (Davami & Sukthankar, 2015); understanding review-to-behavior fidelity in mobile applications (Kong et al., 2015); source code summarization (Nazar et al., 2016) and classification of bug reports (J. Wang et al., 2017).

6.8 Education

Regarding education (Table 6), we found several methods trying to solve the problem of grading in contexts where traditional grading is not possible, due, mainly, to problems of scalability. We refer the reader to (Romero & Ventura, 2017), where different challenges of online education are discussed. Specifically, we found applications about: test grading without answers (Bachrach, Graepel, Minka, & Guiver, 2012); ordinal peer grading (Raman & Joachims, 2014); English grading (Shashidhar, Pandey, & Aggarwal, 2015) and peer grading taking into account both answers and grading (Labutov & Studer, 2017).

6.9 Information Retrieval

In the area of information retrieval (Table 6), there are applications related to learning relevance of medical documents (Wilbur & Kim, 2011), circumlocution of queries (Stanton, Ieong, & Mishra, 2014) and learning topic models (Rodrigues, Ribeiro, Lourenço, & Pereira, 2015).

6.10 Other applications

There are a great number of applications in other domains, such as security or energy, which are also relevant for this study (Table 6): deduplication of digital libraries (Georgescu, Pham, Firan, Nejdl, & Gaugaz, 2012); imitation learning (Chung, Forbes, Cakmak, & Rao, 2014);

evaluation of procedural content generation (Roberts & Chen, 2015); action model acquisition (Zhuo, 2015); weighting antivirus labels (Kantchelian et al., 2015); aerosol optical depth estimation (Djuric, Kansakar, & Vucetic, 2016); point of interest labeling (Hu et al., 2016); detection of spatial events (Ouyang, Srivastava, Toniolo, & Norman, 2016); interstate conflict measurement (D'Orazio, Kenwick, Lane, Palmer, & Reitter, 2016); annotation of energy data (Cao, Rauchenstein, Wijaya, Aberer, & Nunes, 2016); extracting semantic attributes to describe concepts (Tian, Chen, & Zhu, 2017); category learning (Danileiko & Lee, 2017); crowd databases (Robinson, Luo, Sponaugle, Guigand, & Cowen, 2017) and network quality measurements (Y. Li et al., 2017).

Type	Application	Aggregation	Crowd	Platform
Education	(Bachrach et al., 2012)	Complex	Volunteers	Custom
	(Raman & Joachims, 2014)	Complex	Crowd	Custom
	(Shashidhar et al., 2015)	No	Crowd	AMT
	(Labutov & Studer, 2017)	Complex	Crowd	AMT
Information Retrieval	(Wilbur & Kim, 2011)	No	Experts	Custom
	(Stanton et al., 2014)	No	Crowd	CrowdFlower
	(Rodrigues et al., 2015)	Complex	Crowd	AMT
Others	(Georgescu et al., 2012)	Complex	Crowd	AMT
	(Chung et al., 2014)	Simple	Crowd	AMT
	(Roberts & Chen, 2015)	Complex	Experts	Custom
	(Zhuo, 2015)	Complex	Volunteers	Custom
	(Kantchelian et al., 2015)	Complex	Algorithms	Custom
	(Djuric et al., 2016)	Complex	Instruments	Custom
	(Hu et al., 2016)	Complex	Crowd	ChinaCrowds
	(Ouyang et al., 2016)	Complex	Volunteers	Custom
	(D'Orazio et al., 2016)	Simple	Crowd	AMT
	(Cao et al., 2016)	Simple	Crowd	Custom
	(Tian et al., 2017)	Complex	Features	AMT
	(Danileiko & Lee, 2017)	Simple	Labels	Custom
	(Robinson et al., 2017)	Simple	Both	Custom
	(G. Li et al., 2017)	Complex	Crowd	CrowdFlower,ChinaCrowds
	(Y. Li et al., 2017)	Complex	Volunteers	Custom

Table 6: Publications about education, information retrieval and other topics

7 Future research in the field

In this section we comment on future lines of research found in the publications analyzed. To perform this analysis we extracted unsolved problems from the applications and categorized them into meaningful groups in order to facilitate the analysis of their recurrency, which can be seen in Figure 10.

One of the most common needs revealed in the publications was to find some way to model the instance difficulty for a task (Chung et al., 2014; Duan et al., 2015; Nguyen-Dinh et al., 2013; Cao et al., 2016; Wan & Aggarwal, 2014; Ni et al., 2013). Some authors (Chung

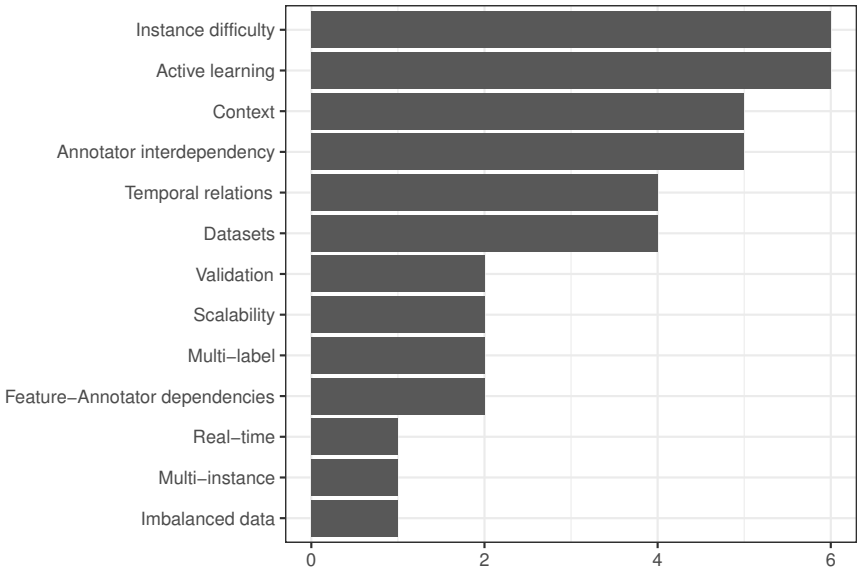


Figure 10: Distribution of publications by proposed problems

et al., 2014) emphasize that the difficulty for a task could be used to reduce the number of annotations required, and hence the cost, for the correct coverage of an example, thus using fewer annotators for easy examples, while collecting more labels for the most difficult ones. Other authors highlight that estimating the difficulty of annotators could improve the reliability of the annotator performance estimation, which could lead to more powerful models (Duan et al., 2015; Nguyen-Dinh et al., 2013; Cao et al., 2016; Wan & Aggarwal, 2014; Ni et al., 2013). There are some classic crowdsourcing algorithms that can take into account instance difficulty, such as (Whitehill, fan Wu, Bergsma, Movellan, & Ruvolo, 2009; Welinder, Branson, Perona, & Belongie, 2010; Donmez, Carbonell, & Schneider, 2009), which could be used directly or as a basis for the development of new algorithms that consider the restrictions of the problem at hand, such as scalability to a large number of examples or the time complexity of the algorithm.

The other most common necessity is the development of active learning techniques (Nguyen-Dinh et al., 2014; S. Chang et al., 2015; Wilbur & Kim, 2011; Nguyen-Dinh et al., 2013; Rodrigues et al., 2014; Salek et al., 2013). As highlighted by the authors, these techniques would not only provide a way to reduce costs when selecting the next example to be annotated, but they could also be used to select which annotator is best for labeling each example, or if the example should be annotated by an expert, according to its difficulty. There are some proposed algorithms (Y. Yan, Fung, Rosales, & Dy, 2011; Zhong, Tang, & Zhou, 2015; J. Zhang, Wu, & Shengs, 2015; M. Fang, Zhu, Li, Ding, & Wu, 2012) that tackle this problem and which may be adapted for particular problems.

The analysis of annotator interdependency is also of interest, as several authors point out

(Kantchelian et al., 2015; Chatelain et al., 2013; Djuric et al., 2016; Ouyang et al., 2016; Zhu, Dunkley, et al., 2015). Most of the models assume that annotators are independent of each other, which is generally not true. The relaxation of these restrictions could be desirable to obtain more powerful models, capable of learning relations between annotators, or even communities within them (Chatelain et al., 2013).

Another necessity to be considered is the adaptation of standard approaches for leveraging temporal relations. Several problems related to temporality have been highlighted by several authors. In (Stanton et al., 2014), for the problem of diagnostic medical queries, the authors propose the exploration of sessions, a sequence of queries about the same topic, although with different search strings. This idea could be applied to annotations, grouping them into sessions for each participant, and analyzing relations between sessions and within the same session. In (D'Orazio et al., 2016), the authors propose the analysis of how the relations between political actors and events change over time. This could also be applied directly to the problem of crowdsourcing, analyzing how the relations between the data and the annotators change with time. In addition, not only the relations between actors, but the evolution of the accuracy and bias of an annotator as time increases could also be studied. This might provide very interesting insights into the learning component of a task and even the loss in accuracy related to boredom or fatigue (Cao et al., 2016; Zhu, Dunkley, et al., 2015).

There is also an increasing need for more datasets with the goal of making the analysis of the performance of new developments as general as possible (Huang et al., 2015; Mitry et al., 2013; Hantke et al., 2016; Fornaciari & Poesio, 2014). As shown in (J. Zhang et al., 2016), there are several public datasets for the problem of learning from crowdsourced data. However, some of them may need non-trivial preprocessing and feature extraction prior to their utilization in algorithm comparison.

Another very interesting proposal is the inclusion of contextual information in the estimation of the reliability of each annotator. Adding information about annotators, such as for example, past experience or age, could be very beneficial when estimating annotator performance (Zhu, Pimentel, et al., 2015). Even, as the authors of (Costa et al., 2011) state, information of the country of origin or mother tongue could be very useful in some problems in which culture plays an important role. Furthermore, other complex relations and information about an annotator could be crowdsourced, capturing high level information that might be of use (Luengo-Oroz et al., 2012). In addition, obtaining data about behaviours when annotators perform a task, such as the time to complete it, could be leveraged to improve annotations and annotator estimations (Cao et al., 2016).

There is also interest in analyzing the scalability of crowdsourcing (Shashidhar et al., 2015) and in adapting algorithms to the MapReduce paradigm (Ouyang et al., 2016). Related

to this, in (Shashidhar et al., 2015) the authors also express the need to develope real-time crowdsourcing algorithms. There is also interest in using crowdsourcing with multi-label (Mavandadi et al., 2012; González et al., 2015) and multi-instance (Tu et al., 2016) problems. Some authors also mention the importance of exploring relations between features and annotators (Ni et al., 2013; Rodrigues et al., 2014).

8 Conclusions

With the rapid growth of the crowdsourcing field, several machine learning applications have appeared that are designed to solve problems that previously may have been unfeasible or to improve the results for problems which use several annotation sources to estimate a ground truth. Furthermore, the new era of big data opens the door to sources which may not be as accurate as those required by traditional machine learning algorithms. This may be the case of data coming from social networks or physical sensors. To tackle these, a different approach, as seen in the applications studied, may be required.

In this paper we have analyzed applications related to learning from crowdsourced labels, indicating the interest in the field as well as several features of these applications, from perspectives such as the type of technique used or the type of crowd employed. We have also analyzed some of the problems that remain unsolved in this field, which may open the door to relevant new research. In particular, we would like to highlight three problems that seem of particular interest and that are not completely solved:

- **Instance difficulty**. Several publications state the need for a way to accurately estimate instance difficulty from crowdsourced data, which may reduce the annotation cost and improve the results. Even though there are some approaches that address this problem, the scalability of the algorithms could be the subject of future research.

- **Annotator interdependency**. Most of the machine learning models proposed for the problem of learning from crowds assume that annotators are independent of each other, which is not usually true. The relaxation of this restriction could lead to learning about communities or groups among annotators with similar characteristics, which could be of interest on its own or useful for improving the results obtained for a problem.

- **Temporal relations**. Another recurrent problem in the applications analyzed is the inexistence of algorithms dealing with temporal relations. This relation could exist between instances, with phenomena appearing such as concept drift or annotators, who could be influenced learning or by getting tired during the process of annotation.

Apart from these problems, we have identified several interesting machine learning problems problems that, when learning from crowds, have not been explored in the literature. In our opinion, that could lead to future research:

- **Supervised feature selection**. There are many applications in which a previous feature selection is of great importance to obtain an adequate machine learning model, as well as to analyze the importance of different characteristics in a problem. To our knowledge, there has not been any effort to adapt the classical algorithms for feature selection to crowdsourced data, which could be very beneficial for the field.

- **Supervised feature discretization**. Several algorithms, as, for example, Bayesian Networks, do not handle continuous features properly. Applying supervised discretization algorithms could make the use of this kind of algorithms easier, as well as improve the interpretability of the analysis.

- **Visualization of crowdsourced data**. In the field of data science, an important task is the exploration and visualization of the data, so that the practitioner is able to decide how to approach further analysis. We believe that, as this kind of data collection is important both for researchers and data scientist alike, correct techniques for visualizing crowsourced data will be undoubtedly helpful.

9 Acknowledgements

This work has been partially funded by the Spanish Research Agency (AEI/MINECO) and FEDER (UE) through project TIN2016-77902-C3-1-P.

References

Aggarwal, J. K., & Ryoo, M. S. (2011). Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, *43*(3), 16.

Albarqouni, S., Baur, C., Achilles, F., Belagiannis, V., Demirci, S., & Navab, N. (2016). Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Transactions on Medical Imaging*, *35*(5), 1313–1321.

Ataer-Cansizoglu, E., Kalpathy-Cramer, J., You, S., Keck, K., Erdogmus, D., Chiang, M., et al. (2015). Analysis of underlying causes of inter-expert disagreement in retinopathy of prematurity diagnosis. *Methods of information in medicine*, *54*(1), 93–102.

Bachrach, Y., Graepel, T., Minka, T., & Guiver, J. (2012). How to grade a test without knowing the answers—a bayesian graphical model for adaptive crowdsourcing and aptitude testing. *Proceedings of the 29th International Conference on Machine Learning*.

Baklanov, A., Fritz, S., Khachay, M., Nurmukhametov, O., & See, L. (2016). The cropland capture game: good annotators versus vote aggregation methods. In *Advanced computational methods for knowledge engineering* (pp. 167–180). Springer.

Bernaschina, C., Fraternali, P., Galli, L., Martinenghi, D., & Tagliasacchi, M. (2014). Robust aggregation of gwap tracks for local image annotation. In *Proceedings of international conference on multimedia retrieval* (p. 403).

Brady, C. J., Mudie, L. I., Wang, X., Guallar, E., & Friedman, D. S. (2017). Improving consensus scoring of crowdsourced data using the rasch model: Development and refinement of a diagnostic instrument. *Journal of medical Internet research*, *19*(6).

Brew, A., Greene, D., & Cunningham, P. (2010). Using crowdsourcing and active learning to track sentiment in online media. In *European conference on artificial intelligence* (pp. 145–150).

Cabezas, F., Carlier, A., Charvillat, V., Salvador, A., & Giro-i Nieto, X. (2015). Quality control in crowdsourced object segmentation. In *Ieee international conference on image processing* (pp. 4243–4247).

Cao, H.-Â., Rauchenstein, F., Wijaya, T. K., Aberer, K., & Nunes, N. (2016). Leveraging user expertise in collaborative systems for annotating energy datasets. In *Ieee international conference on big data* (pp. 3087–3096).

Chang, S., Dai, P., Chen, J., & Chi, E. H. (2015). Got many labels?: Deriving topic labels from multiple sources for social media posts using crowdsourcing and ensemble learning. In *Proceedings of the 24th international conference on world wide web* (pp. 397–406).

Chang, V., Garcia, A., Hitschfeld, N., & Härtel, S. (2017). Gold-standard for computer-assisted morphological sperm analysis. *Computers in Biology and Medicine*, *83*, 143–150.

Chapaneri, S., & Jayaswal, D. (2017). Structured prediction of music mood with twin gaussian processes. In *International conference on pattern recognition and machine intelligence* (pp. 647–654).

Chatelain, P., Pauly, O., Peter, L., Ahmadi, S.-A., Plate, A., Bötzel, K., & Navab, N. (2013). Learning from multiple experts with random forests: Application to the segmentation of the midbrain in 3d ultrasound. In *International conference on medical image computing and computer-assisted intervention* (pp. 230–237).

Chen, B., & Cheng, H. (2010). A review of the applications of agent technology in traffic and transportation systems. *IEEE Transactions on Intelligent Transportation Systems*, *11*(2), 485-497.

Chen, L., Hoey, J., Nugent, C. D., Cook, D. J., & Yu, Z. (2012). Sensor-based activity recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications*

and Reviews), *42*(6), 790–808.

Chesnokova, O., Nowak, M., & Purves, R. S. (2017). A crowdsourced model of landscape preference. In *Lipics-leibniz international proceedings in informatics* (Vol. 86).

Chittaranjan, G., Aran, O., & Gatica-Perez, D. (2011). Exploiting observers' judgements for nonverbal group interaction analysis. In *Ieee international conference on automatic face & gesture recognition* (pp. 734–739).

Chittilappilly, A. I., Chen, L., & Amer-Yahia, S. (2016). A survey of general-purpose crowdsourcing techniques. *IEEE Transactions on Knowledge and Data Engineering*, *28*(9), 2246–2266.

Chung, M. J.-Y., Forbes, M., Cakmak, M., & Rao, R. P. (2014). Accelerating imitation learning through crowdsourcing. In *Ieee international conference on robotics and automation* (pp. 4777–4784).

Costa, J., Silva, C., Antunes, M., & Ribeiro, B. (2011). Get your jokes right: ask the crowd. In *International conference on model and data engineering* (pp. 178–185).

Crescenzi, V., Merialdo, P., & Qiu, D. (2013). Wrapper generation supervised by a noisy crowd. In *Vldb workshop on databases and crowdsourcing* (pp. 8–13).

Danileiko, I., & Lee, M. D. (2017). A model-based approach to the wisdom of the crowd in category learning. *Cognitive science*.

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2006). Studying aesthetics in photographic images using a computational approach. In *European conference on computer vision* (pp. 288–301).

Davami, E., & Sukthankar, G. (2015). Improving the performance of mobile phone crowdsourcing applications. In *Proceedings of the 2015 international conference on autonomous agents and multiagent systems* (pp. 145–153).

de Bruijne, M. (2016). *Machine learning approaches in medical image analysis: From detection to diagnosis.* Elsevier.

DeFelipe, J., López-Cruz, P. L., Benavides-Piccione, R., Bielza, C., Larrañaga, P., Anderson, S., ... others (2013). New insights into the classification and nomenclature of cortical gabaergic interneurons. *Nature Reviews Neuroscience*, *14*(3), 202–216.

de Herrera, A. G. S., Schaer, R., Antani, S., & Müller, H. (2016). Using crowdsourcing for multi-label biomedical compound figure annotation. In *International workshop on large-scale annotation of biomedical data and expert label synthesis* (pp. 228–237).

Djuric, N., Kansakar, L., & Vucetic, S. (2016). Semi-supervised combination of experts for aerosol optical depth estimation. *Artificial Intelligence*, *230*, 1–13.

Donmez, P., Carbonell, J. G., & Schneider, J. (2009). Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th acm sigkdd international conference on knowledge discovery and data mining* (pp. 259–268).

D'Orazio, V., Kenwick, M., Lane, M., Palmer, G., & Reitter, D. (2016). Crowdsourcing the measurement of interstate conflict. *PloS one*, *11*(6).

Duan, L., Oyama, S., Sato, H., & Kurihara, M. (2015). Multi-emotion estimation in narratives from crowdsourced annotations. In *Proceedings of the 15th acm/ieee-cs joint conference on digital libraries* (pp. 91–100).

Elsevier. (2017). *Scimago journal & country rank.* http://www.scimagojr.com/. (Accessed: 2017-05-15)

Fang, M., Zhu, X., Li, B., Ding, W., & Wu, X. (2012). Self-taught active learning from crowds. In *12th international conference on data mining* (pp. 858–863).

Fang, Y.-L., Sun, H.-L., Chen, P.-P., & Deng, T. (2017). Improving the quality of crowdsourced image labeling via label similarity. *Journal of Computer Science and Technology*, *32*(5), 877–889.

Fleuren, W. W., & Alkema, W. (2015). Application of text mining in the biomedical domain. *Methods*, *74*, 97–106.

Foody, G. M., See, L., Fritz, S., Van der Velde, M., Perger, C., Schill, C., ... Comber, A. (2015). Accurate attribute mapping from volunteered geographic information: issues of volunteer quantity and quality. *The Cartographic Journal*, *52*(4), 336–344.

Fornaciari, T., & Poesio, M. (2014). Identifying fake amazon reviews as learning from crowds. In *Proceedings of the 14th conference of the european chapter of the association for computational linguistics.*

Gao, H., Liu, C. H., Wang, W., Zhao, J., Song, Z., Su, X., ... Leung, K. K. (2015). A survey of incentive mechanisms for participatory sensing. *IEEE Communications Surveys & Tutorials*, *17*(2), 918–943.

Garcia-Molina, H., Joglekar, M., Marcus, A., Parameswaran, A., & Verroios, V. (2016). Challenges in data crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering*, *28*(4), 901–911.

Garcia-Ulloa, D. A., Xiong, L., & Sunderam, V. (2017). Truth discovery for spatio-temporal events from crowdsourced data. *Proceedings of the VLDB Endowment*, *10*(11), 1562–1573.

Georgescu, M., Pham, D. D., Firan, C. S., Nejdl, W., & Gaugaz, J. (2012). Map to humans and reduce error: crowdsourcing for deduplication applied to digital libraries. In *Proceedings of the 21st acm international conference on information and knowledge management* (pp. 1970–1974).

González, J. G., Álvarez, M. A., & Orozco, Á. A. (2015). Automatic assessment of voice quality in the context of multiple annotations. In *37th annual international conference of the ieee on engineering in medicine and biology society* (pp. 6236–6239).

Good, B. M., & Su, A. I. (2013). Crowdsourcing for bioinformatics. *Bioinformatics*, btt333.

Greenwood, M., Elwyn, G., Francis, N., Preece, A., & Spasic, I. (2013). Automatic extraction of personal experiences from patients' blogs: A case study in chronic obstructive pulmonary disease. In *Third international conference on cloud and green computing* (pp. 377–382).

Gueguen, L., Koenig, J., Reeder, C., Barksdale, T., Saints, J., Stamatiou, K., ... Johnston, C. (2017). Mapping human settlements and population at country scale from vhr images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *10*(2), 524-538.

Hantke, S., Marchi, E., & Schuller, B. (2016). Introducing the weighted trustability evaluator for crowdsourcing exemplified by speaker likability classification. In *Proceedings 10th language resources and evaluation conference.*

Hantke, S., Zhang, Z., & Schuller, B. (2017). Towards intelligent crowdsourcing for audio data annotation: integrating active learning in the real world. In *Proceedings interspeech.*

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261–266.

Hovy, D., Plank, B., & Søgaard, A. (2014). Experiments with crowdsourced re-annotation of a pos tagging data set. In *Annual meeting of the association for computational linguistics* (pp. 377–382).

Howe, J. (2006). The rise of crowdsourcing. *Wired magazine*, *14*(6), 1–4.

Hu, H., Zheng, Y., Bao, Z., Li, G., Feng, J., & Cheng, R. (2016). Crowdsourced poi labelling: Location-aware result inference and task assignment. In *Ieee 32nd international conference on data engineering* (pp. 61–72).

Huang, Z., Zhong, J., & Passonneau, R. J. (2015). Estimation of discourse segmentation labels from crowd data. In *Conference on empirical methods in natural language processing* (pp. 2190–2200).

Jain, S., Kashyap, R., Kuo, T.-T., Bhargava, S., Lin, G., & Hsu, C.-N. (2016). Weakly supervised learning of biomedical information extraction from curated data. In *Bmc bioinformatics* (Vol. 17, p. 1).

Jia, X., Khandelwal, A., Gerber, J., Carlson, K., West, P., & Kumar, V. (2016). Learning large-scale plantation mapping from imperfect annotators. In *Big data (big data), 2016 ieee international conference on* (pp. 1192–1201).

Jones, G. J. (2012). An introduction to crowdsourcing for language and multimedia technology research. In *Promise winter school on information retrieval meets information visualization* (pp. 132–154).

Kantchelian, A., Tschantz, M. C., Afroz, S., Miller, B., Shankar, V., Bachwani, R., ... Tygar, J. D. (2015). Better malware ground truth: Techniques for weighting anti-virus vendor labels. In *Proceedings of the 8th acm workshop on artificial intelligence*

and security (pp. 45–56).

Kaster, F. O., Menze, B. H., Weber, M.-A., & Hamprecht, F. A. (2010). Comparative validation of graphical models for learning tumor segmentations from noisy manual annotations. In *International miccai workshop on medical computer vision* (pp. 74–85).

Katsimerou, C., Albeda, J., Huldtgren, A., Heynderickx, I., & Redi, J. A. (2016). Crowdsourcing empathetic intelligence: the case of the annotation of emma database for emotion and mood recognition. *ACM Transactions on Intelligent Systems and Technology*, *7*(4), 51.

Kaya, A., & Can, A. B. (2015). A weighted rule based method for predicting malignancy of pulmonary nodules by nodule characteristics. *Journal of biomedical informatics*, *56*, 69–79.

Khare, R., Burger, J. D., Aberdeen, J. S., Tresner-Kirsch, D. W., Corrales, T. J., Hirchman, L., & Lu, Z. (2015). Scaling drug indication curation through crowdsourcing. *Database*.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, *33*(2004), 1–26.

Kołakowska, A., Landowska, A., Szwoch, M., Szwoch, W., & Wrobel, M. R. (2014). Emotion recognition and its applications. In *Human-computer systems interaction: Backgrounds and applications 3* (pp. 51–62). Springer.

Kong, D., Cen, L., & Jin, H. (2015). Autoreb: Automatically understanding the review-to-behavior fidelity in android applications. In *Proceedings of the 22nd acm sigsac conference on computer and communications security* (pp. 530–541).

Kratz, S. G., & Wiese, J. (2016). Gestureseg: developing a gesture segmentation system using gesture execution phase labeling by crowd workers. In *The 8th acm sigchi symposium on engineering interactive computing systems* (pp. 61–72).

Labutov, I., & Studer, C. (2017). Jag: A crowdsourcing framework for joint assessment and peer grading. In *Aaai* (pp. 1010–1016).

Li, G., Chai, C., Fan, J., Weng, X., Li, J., Zheng, Y., ... Yuan, H. (2017). Cdb: optimizing queries with crowd-based selections and joins. In *Proceedings of the 2017 acm international conference on management of data* (pp. 1463–1478).

Li, G., Wang, J., Zheng, Y., & Franklin, M. J. (2016). Crowdsourced data management: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *28*(9), 2296–2319.

Li, W., Wu, W.-j., Wang, H.-m., Cheng, X.-q., Chen, H.-j., Zhou, Z.-h., & Ding, R. (2017). Crowd intelligence in ai 2.0 era. *Frontiers of Information Technology & Electronic Engineering*, *18*(1), 15–43.

Li, Y., Gao, J., Lee, P. P., Su, L., He, C., He, C., ... Fan, W. (2017). A weighted crowdsourcing approach for network quality measurement in cellular data networks.

IEEE Transactions on Mobile Computing, *16*(2), 300–313.

Li, Z.-X., Yang, Q., Liu, A., Liu, G.-F., Zhu, J., Xu, J.-J., ... Zhang, M. (2017). Crowd-guided entity matching with consolidated textual data. *Journal of Computer Science and Technology*, *32*(5), 858–876.

Lou, Y., Tu, S. W., Nyulas, C., Tudorache, T., Chalmers, R. J., & Musen, M. A. (2017). Use of ontology structure and bayesian models to aid the crowdsourcing of icd-11 sanctioning rules. *Journal of Biomedical Informatics*.

Lu, Z., Kao, H.-Y., Wei, C.-H., Huang, M., Liu, J., Kuo, C.-J., ... others (2011). The gene normalization task in biocreative iii. *BMC bioinformatics*, *12*(8), S2.

Luengo-Oroz, M. A., Arranz, A., & Frean, J. (2012). Crowdsourcing malaria parasite quantification: an online game for analyzing images of infected thick blood smears. *Journal of medical Internet research*, *14*(6).

Ma, F., Meng, C., Xiao, H., Li, Q., Gao, J., Su, L., & Zhang, A. (2017). Unsupervised discovery of drug side-effects from heterogeneous data sources. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 967–976).

Machedon, R., Rand, W., & Joshi, Y. (2013). Automatic crowdsourcing-based classification of marketing messaging on twitter. In *International conference on social computing* (pp. 975–978).

Mahapatra, D., Schüffler, P. J., Tielbeek, J. A., Puylaert, C. A., Makanyanga, J. C., Menys, A., ... others (2014). Combining multiple expert annotations using semi-supervised learning and graph cuts for crohnś disease segmentation. In *International miccai workshop on computational and clinical challenges in abdominal imaging* (pp. 139–147).

Mao, K., Capra, L., Harman, M., & Jia, Y. (2017). A survey of the use of crowdsourcing in software engineering. *Journal of Systems and Software*, *126*, 57–84.

Mavandadi, S., Feng, S., Yu, F., Dimitrov, S., Nielsen-Saines, K., Prescott, W. R., & Ozcan, A. (2012). A mathematical framework for combining decisions of multiple experts toward accurate and remote diagnosis of malaria using tele-microscopy. *PloS one*, *7*(10), e46192.

Mihaljević, B., Benavides-Piccione, R., Guerra, L., DeFelipe, J., Larrañaga, P., & Bielza, C. (2015). Classifying gabaergic interneurons with semi-supervised projected model-based clustering. *Artificial intelligence in medicine*, *65*(1), 49–59.

Min, X., Shi, Y., Cui, L., Yu, H., & Miao, Y. (2017). Efficient crowd-powered active learning for reliable review evaluation. In *Proceedings of the 2nd international conference on crowd science and engineering* (pp. 136–143).

Mitry, D., Peto, T., Hayat, S., Morgan, J. E., Khaw, K.-T., & Foster, P. J. (2013). Crowdsourcing as a novel technique for retinal fundus photography classification: Analysis of

images in the epic norfolk cohort on behalf of the ukbiobank eye and vision consortium. *PloS one*, *8*(8).

Mok, R. K., Chang, R. K., & Li, W. (2017). Detecting low-quality workers in qoe crowdtesting: A worker behavior-based approach. *IEEE Transactions on Multimedia*, *19*(3), 530–543.

Muller, C., Chapman, L., Johnston, S., Kidd, C., Illingworth, S., Foody, G., ... Leigh, R. (2015). Crowdsourcing for climate and atmospheric sciences: Current status and future potential. *International Journal of Climatology*, *35*(11), 3185–3203.

Nazábal, A., García-Moreno, P., Artés-Rodríguez, A., & Ghahramani, Z. (2016). Human activity recognition by combining a small number of classifiers. *IEEE Journal of Biomedical and Health Informatics*, *20*(5), 1342–1351.

Nazar, N., Jiang, H., Gao, G., Zhang, T., Li, X., & Ren, Z. (2016). Source code fragment summarization with small-scale crowdsourcing based features. *Frontiers of Computer Science*, *10*(3), 504–517.

Ng, J.-P., & Kan, M.-Y. (2012). Improved temporal relation classification using dependency parses and selective crowdsourced annotations. In *International conference on computational linguistics* (pp. 2109–2124).

Nguyen, A. T., Wallace, B. C., Li, J. J., Nenkova, A., & Lease, M. (2017). Aggregating and predicting sequence labels from crowd annotations. In *Proceedings of the conference. association for computational linguistics. meeting* (Vol. 2017, p. 299).

Nguyen, Q. V. H., Duong, C. T., Nguyen, T. T., Weidlich, M., Aberer, K., Yin, H., & Zhou, X. (2017). Argument discovery via crowdsourcing. *The VLDB Journal*, *26*(4), 511–535.

Nguyen-Dinh, L.-V., Calatroni, A., & Tröster, G. (2014). Robust online gesture recognition with crowdsourced annotations. *Journal of Machine Learning Research*, *15*, 3187–3220.

Nguyen-Dinh, L.-V., Waldburger, C., Roggen, D., & Tröster, G. (2013). Tagging human activities in video by crowdsourcing. In *Proceedings of the 3rd acm international conference on multimedia retrieval* (pp. 263–270).

Ni, Y., McVicar, M., Santos-Rodriguez, R., & De Bie, T. (2013). Understanding effects of subjectivity in measuring chord estimation accuracy. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(12), 2607–2615.

Nicolaou, M. A., Pavlovic, V., & Pantic, M. (2014). Dynamic probabilistic cca for analysis of affective behavior and fusion of continuous annotations. *IEEE transactions on pattern analysis and machine intelligence*, *36*(7), 1299–1311.

Oosterman, J., Yang, J., Bozzon, A., Aroyo, L., & Houben, G.-J. (2015). On the impact of knowledge extraction and aggregation on crowdsourced annotation of visual artworks.

Computer Networks, *90*, 133–149.

Ouyang, R. W., Srivastava, M., Toniolo, A., & Norman, T. J. (2016). Truth discovery in crowdsourced detection of spatial events. *IEEE Transactions on Knowledge and Data Engineering*, *28*(4), 1047–1060.

Paliwal, M., & Kumar, U. (2009). Neural networks and statistical techniques: A review of applications. *Expert Systems with Applications*, *36*(1), 2-17.

Passonneau, R. J., Bhardwaj, V., Salleb-Aouissi, A., & Ide, N. (2012). Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, *46*(2), 219–252.

Peng, J., Liu, Q., Ihler, A., & Berger, B. (2013). Crowdsourcing for structured labeling with applications to protein folding. In *Icml workshop: Machine learning meets crowdsourcing.*

Pistorius, T., & Poona, N. (2014). Accuracy assessment of game-based crowdsourced land-use/land cover image classification. In *Ieee international geoscience and remote sensing symposium* (pp. 4780–4783).

Raman, K., & Joachims, T. (2014). Methods for ordinal peer grading. In *Proceedings of the 20th acm sigkdd international conference on knowledge discovery and data mining* (pp. 1037–1046).

Rashid, B., & Rehmani, M. H. (2016). Applications of wireless sensor networks for urban areas: a survey. *Journal of Network and Computer Applications*, *60*, 192–219.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., & Moy, L. (2010). Learning from crowds. *Journal of Machine Learning Research*, *11*(Apr), 1297–1322.

Roberts, J., & Chen, K. (2015). Learning-based procedural content generation. *IEEE Transactions on Computational Intelligence and AI in Games*, *7*(1), 88–101.

Robinson, K. L., Luo, J. Y., Sponaugle, S., Guigand, C., & Cowen, R. K. (2017). A tale of two crowds: Public engagement in plankton classification. *Frontiers in Marine Science*, *4*, 82.

Rodrigues, F., Lourenco, M., Ribeiro, B., & Pereira, F. C. (2017). Learning supervised topic models for classification and regression from crowds. *IEEE transactions on pattern analysis and machine intelligence*, *39*(12), 2409–2422.

Rodrigues, F., Pereira, F., & Ribeiro, B. (2014). Sequence labeling with multiple annotators. *Machine Learning*, *95*(2), 165–181.

Rodrigues, F., Ribeiro, B., Lourenço, M., & Pereira, F. (2015). Learning supervised topic models from crowds. In *The aaai conference on artificial intelligence.*

Romero, C., & Ventura, S. (2017). Educational data science in massive open online courses. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*(1).

Rudinac, S., Larson, M., & Hanjalic, A. (2013). Learning crowdsourced user preferences for

visual summarization of image collections. *IEEE Transactions on Multimedia*, *15*(6), 1231–1243.

Salek, M., Bachrach, Y., & Key, P. (2013). Hotspotting-a probabilistic graphical model for image object localization through crowdsourcing. In *The aaai conference on artificial intelligence.*

Salter-Townshend, M., & Murphy, T. B. (2013). Sentiment analysis of online media. In *Algorithms from and for nature and life* (pp. 137–145). Springer.

Sameki, M., Gurari, D., & Betke, M. (2016). Icord: Intelligent collection of redundant dataa dynamic system for crowdsourcing cell segmentations accurately and efficiently. In *Ieee conference on computer vision and pattern recognition workshops* (pp. 1380–1389).

Servajean, M., Joly, A., Shasha, D., Champ, J., & Pacitti, E. (2017). Crowdsourcing thousands of specialized labels: a bayesian active training approach. *IEEE Transactions on Multimedia*, *19*(6), 1376–1391.

Sharma, M., Saha, O., Sriraman, A., Hebbalaguppe, R., Vig, L., & Karande, S. (2017). Crowdsourcing for chromosome segmentation and deep classification. In *Computer vision and pattern recognition workshops (cvprw), 2017 ieee conference on* (pp. 786–793).

Shashidhar, V., Pandey, N., & Aggarwal, V. (2015). Spoken english grading: Machine learning with crowd intelligence. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining* (pp. 2089–2097).

Siddharthan, A., Lambin, C., Robinson, A.-M., Sharma, N., Comont, R., O'mahony, E., ... Wal, R. V. D. (2016). Crowdsourcing without a crowd: Reliable online species identification using bayesian models to minimize crowd size. *ACM Transactions on Intelligent Systems and Technology*, *7*(4), 45.

Silva, I., Behar, J., Sameni, R., Zhu, T., Oster, J., Clifford, G. D., & Moody, G. B. (2013). Noninvasive fetal ecg: the physionet/computing in cardiology challenge 2013. In *Computing in cardiology conference* (pp. 149–152).

Srivastava, G., Yoder, J. A., Park, J., & Kak, A. C. (2013). Using objective ground-truth labels created by multiple annotators for improved video classification: A comparative study. *Computer Vision and Image Understanding*, *117*(10), 1384–1399.

Stanton, I., Ieong, S., & Mishra, N. (2014). Circumlocution in diagnostic medical queries. In *Proceedings of the 37th international acm sigir conference on research & development in information retrieval* (pp. 133–142).

Su, H., Deng, J., & Fei-Fei, L. (2012). Crowdsourcing annotations for visual object detection. In *Workshops at the twenty-sixth aaai conference on artificial intelligence* (Vol. 1).

Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & de Alfaro, L. (2017). Some like it hoax: Automated fake news detection in social networks. In *Ceur workshop*

proceedings (Vol. 1960).

Tan, D., Zhao, R., Sun, J., & Qin, W. (2015). Sleep spindle detection using deep learning: A validation study based on crowdsourcing. In *37th annual international conference of the ieee in engineering in medicine and biology society* (pp. 2828–2831).

Tastan, O., Qi, Y., Carbonell, J. G., & Klein-Seetharaman, J. (2014). Refining literature curated protein interactions using expert opinions. In *Pacific symposium on biocomputing* (pp. 318–329).

Tavares, G., Mourão, A., & Magalhães, J. (2016). Crowdsourcing facial expressions for affective-interaction. *Computer Vision and Image Understanding*, *147*, 102–113.

Tian, T., Chen, N., & Zhu, J. (2017). Learning attributes from the crowdsourced relative labels. In *Aaai* (Vol. 1, p. 2).

Tu, M., Jiao, Y., Berisha, V., & Liss, J. M. (2016). Models for objective evaluation of dysarthric speech from data annotated by multiple listeners. In *50th asilomar conference on signals, systems and computers* (pp. 827–830).

Vijayanarasimhan, S., & Grauman, K. (2014). Large-scale live active learning: Training object detectors with crawled data and crowds. *International Journal of Computer Vision*, *108*(1-2), 97–114.

Wallace, B. C., Noel-Storr, A., Marshall, I. J., Cohen, A. M., Smalheiser, N. R., & Thomas, J. (2017). Identifying reports of randomized controlled trials (rcts) via a hybrid machine learning and crowdsourcing approach. *Journal of the American Medical Informatics Association*, *24*(6), 1165–1168.

Wan, S., & Aggarwal, J. (2014). Spontaneous facial expression recognition: A robust metric learning approach. *Pattern Recognition*, *47*(5), 1859–1868.

Wang, J., Cui, Q., Wang, S., & Wang, Q. (2017). Domain adaptation for test report classification in crowdsourced testing. In *Proceedings of the 39th international conference on software engineering: Software engineering in practice track* (pp. 83–92).

Wang, S., & Summers, R. M. (2012). Machine learning and radiology. *Medical image analysis*, *16*(5), 933–951.

Welinder, P., Branson, S., Perona, P., & Belongie, S. J. (2010). The multidimensional wisdom of crowds. In *Advances in neural information processing systems* (pp. 2424–2432).

Whitehill, J., fan Wu, T., Bergsma, J., Movellan, J. R., & Ruvolo, P. L. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems 22* (pp. 2035–2043).

Wilbur, W. J., & Kim, W. (2011). Improving a gold standard: treating human relevance judgments of medline document pairs. *BMC bioinformatics*, *12*(3), S5.

Wu, C.-C., Chen, K.-T., Chang, Y.-C., & Lei, C.-L. (2013). Crowdsourcing multimedia qoe evaluation: A trusted framework. *IEEE transactions on multimedia*, *15*(5), 1121–

1137.

Yan, R., Song, Y., Li, C.-T., Zhang, M., & Hu, X. (2015). Opportunities or risks to reduce labor in crowdsourcing translation? characterizing cost versus quality via a pagerank-hits hybrid model. In *International joint conference on artificial intelligence* (pp. 1025–1032).

Yan, Y., Fung, G. M., Rosales, R., & Dy, J. G. (2011). Active learning from crowds. In *Proceedings of the 28th international conference on machine learning* (pp. 1161–1168).

Zhang, J., Wu, X., & Sheng, V. S. (2016). Learning from crowdsourced labeled data: a survey. *Artificial Intelligence Review*, *46*(4), 543–576.

Zhang, J., Wu, X., & Shengs, V. S. (2015). Active learning with imbalanced multiple noisy labeling. *IEEE transactions on cybernetics*, *45*(5), 1095–1107.

Zhang, S., Vempaty, A., Parks, S. E., & Varshney, P. K. (2017). On classification of environmental acoustic data using crowds. In *Acoustics, speech and signal processing (icassp), 2017 ieee international conference on* (pp. 5880–5884).

Zhang, X., Yang, Y.-H., Han, Z., Wang, H., & Gao, C. (2013). Object class detection: A survey. *ACM Computing Surveys*, *46*(1). (cited By 20) doi: 10.1145/2522968.2522978

Zheng, Y., Li, G., Li, Y., Shan, C., & Cheng, R. (2017). Truth inference in crowdsourcing: is the problem solved? *Proceedings of the VLDB Endowment*, *10*(5), 541–552.

Zhong, J., Tang, K., & Zhou, Z.-H. (2015). Active learning from crowds with unsure option. In *International joint conference on artificial intelligence* (pp. 1061–1068).

Zhou, L.-k., Tang, S.-l., Xiao, J., Wu, F., & Zhuang, Y.-t. (2017). Disambiguating named entities with deep supervised learning via crowd labels. *Frontiers of Information Technology & Electronic Engineering*, *18*(1), 97–106.

Zhu, T., Dunkley, N., Behar, J., Clifton, D. A., & Clifford, G. D. (2015). Fusing continuous-valued medical labels using a bayesian model. *Annals of biomedical engineering*, *43*(12), 2892–2902.

Zhu, T., Johnson, A. E., Behar, J., & Clifford, G. D. (2014). Crowd-sourced annotation of ecg signals using contextual information. *Annals of biomedical engineering*, *42*(4), 871–884.

Zhu, T., Pimentel, M. A., Clifford, G. D., & Clifton, D. A. (2015). Bayesian fusion of algorithms for the robust estimation of respiratory rate from the photoplethysmogram. In *37th annual international conference of the ieee on engineering in medicine and biology society* (pp. 6138–6141).

Zhuo, H. H. (2015). Crowdsourced action-model acquisition for planning. In *The aaai conference on artificial intelligence* (pp. 3439–3446).

A Search strings

A.1 Scopus

The search string used in Scopus was:

```
TITLE(
(learning or classification or model or
inference or supervised)
AND
(crowds or crowdsourced or crowdsourcing or
annotators or labelers))
AND
( LIMIT-TO(DOCTYPE,"cp" ) OR
LIMIT-TO(DOCTYPE,"ar" ) )
AND
(EXCLUDE(EXACTKEYWORD,"Crowd Simulation"))
AND
(EXCLUDE(SRCTYPE,"k" ) OR
EXCLUDE(SRCTYPE,"d"))
AND
(EXCLUDE(EXACTKEYWORD,"Computer Simulation"))
```

With this string, apart from searching for the terms related to machine learning and crowdsourcing, we exclude some areas that are of no interest for our research, namely the areas related to simulation. We also limit the search for publications to journals and conferences.

A.2 Web Of Science

The search string used in Web Of Science was:

```
(TI=(learning OR classification OR
model OR inference OR supervised)
AND
TI=(crowds OR crowdsourced OR
crowdsourcing OR annotators OR
labellers))
NOT TI=(pedestrians OR dynamics OR
segregation OR lanes OR crowding)
```

As in the previous case, we limit the search to the terms related to learning from crowdsourced data and exclude several terms related to crowd prediction or crowding.

A.3 Google Scholar

The search string used in Google Scholar was:

```
"(learning OR classification OR
inference OR supervised)
AND
(crowd OR crowdsourced OR
crowdsourcing OR annotations OR
labelers)"
```

The meaning of this string is similar to the one above. We search for at least one of the terms inside the parenthesis. The title should have at least one term from each group.

A.4 DBLP

The search string used in DBLP was:

```
(learning|classification|
inference|supervised)
(crowds|crowdsourced|crowdsourcing|
annotators|labelers)
```

A.5 ScienceDirect

The search string used in ScienceDirect was:

```
ttl((learning or classification or
inference or supervised)
AND (crowds or crowdsourced or
crowdsourcing or annotators
or labelers))
```

A.6 ACM Digital Library

The search string used in the ACM Digital Library was:

```
learning OR classification OR
inference OR supervised
AND crowds OR crowdsourced OR
crowdsourcing OR annotators OR
labelers
```

A.7 IEEE Xplore Digital Library

The search string used in the IEEE Xplore Digital Library was:

```
(("Document Title":"crowdsourcing" OR
"Document Title":"crowds" OR
"Document Title":"crowdsourced" OR
"Document Title":"annotators" OR
"Document Title":"labelers")
AND
(p_Title:"learning" OR
"Document Title":"classification" OR
"Document Title":"model" OR
"Document Title":"inference" OR
"Document Title":"supervised"))
```